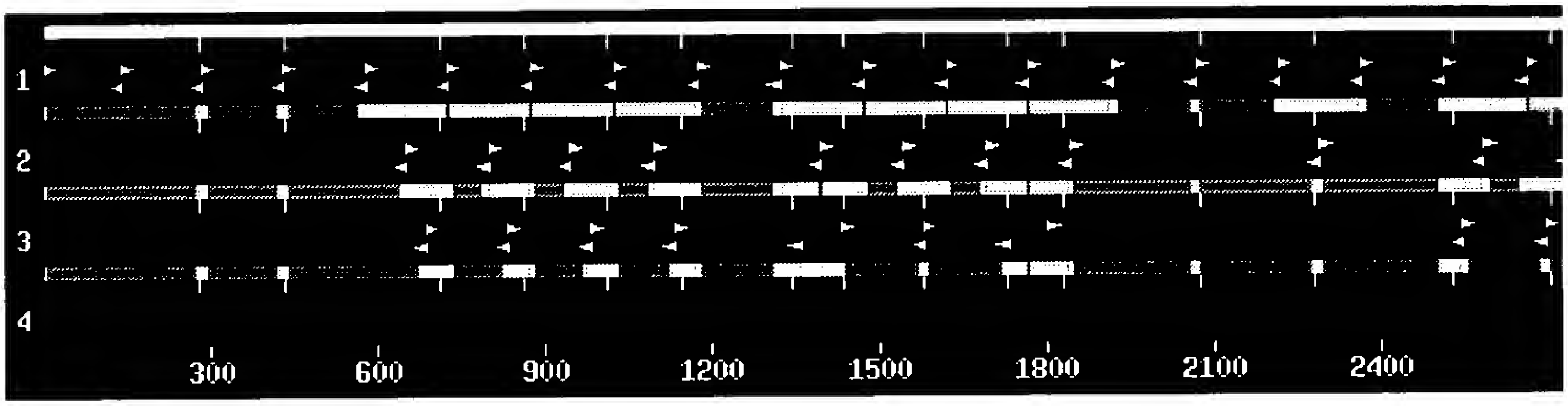
ExonPCR:
Exon Detection in cDNA with PCR Experiments

Under construction



This software implements the first step of the ExonPCR protocol for finding the approximate locations of exon boundaries in a cDNA using results from multiple PCR experiments described in Xu G., Sze S.-H., Liu C.-P., Pevzner P.A. and Arnheim N. (1998) Gene hunting without sequencing genomic clones: finding exon boundaries in cDNAs. *Genomics*, **47**, 171-179. Primers are designed from the cDNA sequence and used to amplify genomic DNA. Each pair of primers serves as a query asking the question whether, in genomic DNA, there exists an intron(s) between the primer sequences. The answer to this query is provided by comparison of the length of PCR products in cDNA and genomic DNA. We recognize three answers to this query. A pair of primers located within the same exon will produce the same-sized PCR product with both templates (+ result). If the PCR product using genomic DNA is longer than expected from the cDNA sequence or is missing altogether, the region is interrupted by an intron(s) in genomic DNA (- result). Results the status of which cannot be defined as either + or - due to experimental uncertainties are called ? results.

A carefully devised strategy that minimizes the total number of PCR primers that are used (to reduce cost) and at the same time minimizes the total number of required rounds of PCR experiments (to reduce time) would be of great value. The goals of minimizing both the total number of primers and the number of rounds conflict with each other. A minimum number of primer pairs is achieved in a sequential dichotomy protocol in which only one primer pair is designed in every round based on the results of earlier rounds of experiments. This strategy is unrealistic since it leads to an excessive number of rounds. An alternative single-round protocol designs all possible primer pairs in a single round thus leading to an excessively large number of primers. Since these criteria are conflicting, we search for a tradeoff between the dichotomy strategy and the single round strategy. The goal is not to determine exon boundaries exactly, but to find the approximate locations of exon boundaries to within a small resolution (say 30 to 100 bp). The second step of the ExonPCR protocol then uses ligation-mediated PCR to determine the exon boundaries with the help of limited DNA sequencing (not implemented in this software).

Input to each round includes the cDNA sequence, primers used in previous rounds and experimental results. The software designs primers for the new round.

The software does not check for primer or primer pair stability extensively. Given minimum and maximum primer lengths, the software returns primers that have expected melting

temperatures within a legal range based on a formula proposed in Wu et al. (1991) *DNA and Cell Biology*, 10, 233-238. The software doesn't check if primers have significant self overlaps or if primer pairs have significant overlaps. Slight adjustments of primer positions (either to the left or to the right) may be necessary to get stable primer pairs. Primer stability can be checked extensively with software such as xprimer.

⚠ **The software requires saving plain text files in the local system between rounds by using the file saving utility of the web browser. However, some software programs on Mac or PC don't handle newlines from the web browser correctly, although they can handle plain text in general. These software programs can still be used for saving, though, by first creating an empty file from within the software, doing copy and paste of the entire plain text display to the empty file, and then saving the file.**

⚠ **In some cases, the web browser has to be configured so that it displays a plain text file directly instead of prompting the user. The general procedure is to associate the file type text/plain with the default suffix txt and have it handled directly by the browser.**

Software for adaptive exon detection in cDNA

Example

Input | Algorithm | Output | Recommendations

Input

- **Sequence**. Normally, it is the given cDNA sequence. The sequence can be named for reporting purpose. All upper-case and lower-case characters are recognized. In test mode, it is the unspliced genomic sequence and its sequence structure should be specified so that the cDNA sequence can be extracted.
- **Sequence structure (test mode)**. The gene structure can be specified as pairs of numbers indicating exons. Characters other than digits are ignored. This is used only when the exon boundaries are known in test mode to determine the cDNA sequence.
- **Last round**. This field is for the specification of results in previous rounds. Normally, for each round, the software will generate a plain file to be used for the new round by listing the set of new primers and primer pairs. The user decides what experiments to perform and update this file with experimental results. The file can then be pasted into this field to start the next round. However, this field has a fixed format (which will be described) and the user is free to use any input that fits this format, perhaps adding results from other sources (other than from PCR experiments) to the field.

  Each line of the field forms a command. Each command consists of words separated by spaces. There are four types of words (no spaces within words):
  1. **primer_name**. Each primer name is formed from three parts (with no spaces in between): first the round number in which the primer is generated, second the character > or < (> indicates left primer, < indicates right primer), and finally a string naming the primer.
  2. **position**. A position is specified by a pair of numbers in the format n1..n2, where n1

specifies the start and n2 specifies the end.
3. **round_number**. A positive number by its own specifies a round number.
4. **type**. It is either +, - or ?. The meanings vary slightly in different contexts. In general, + means there are no exon boundaries, - means there is at least one exon boundary, and ? means unknown status.

The three types of commands are as follows (square brackets indicate optional elements, <space> stands for at least one space):

1. **Primer definition (syntax: primer_name[type] <space> position)**. This command defines a primer. A primer name should not be redefined later and a primer cannot occupy exactly the same location as another primer. An optional type specification immediately following the primer name (with no spaces in between) gives the strongest type of experimental result that the primer is involved. In other words, type is + if there is a + result with the primer, type is - if there is no + result but there is a - result with the primer, and type is ? if the primer is involved only in ? results. This helps the user to quickly see the strongest result type associated with a primer without going through the possibly large set of experimental results associated with the primer.
2. **Experimental result (syntax: round_number <space> left_primer_name <space> right_primer_name <space> [type])**. The command allows the user to specify an experimental result. The type specification is optional and the command has no effect if it is not present. This feature allows the program to process partially filled results. It is very helpful when the number of results to be filled is very large and the user wants to check the current status with partial results.
3. **Region type (syntax: position <space> type)**. This command specifies the type of a region. Regions can be overlapping.

The program processes each line sequentially, checking both for syntax errors and inconsistencies to previous results. An error in a line invalidates the line, that is, the line has no effect (one exception is that extra characters in a line are ignored and the line has effect). The program ignores all lines with errors and proceeds even in the presence of errors. The user can also choose whether results should be interpreted in a conservative way or not. Either approach interprets - results as spanning the outer ends of a primer pair, observing the possibility that an exon boundary can occur inside either primer of the pair. The conservative approach interprets + results as spanning up to the midpoints of both primers in the pair, while the non-conservative approach interprets + results as spanning the outer ends of a primer pair.

- **Minimum melting temperature**.
- **Maximum melting temperature**.
- **Minimum length of primers**.
- **Maximum length of primers**.
- **Minimum distance between primer ends**. It is the minimum distance between the outer ends of a primer pair to be used in PCR. Its value has to be at least twice the minimum primer length.
- **Maximum distance between primer ends**.
- **Resolution**. This is the desirable maximum size of - or ? regions after the new round. While the algorithm will try to generate primers so that the specified resolution can be satisfied after the new round, it will not always be successful (in some cases, it is theoretically impossible). There is no guarantee that the resolution will be satisfied after the new round. Its value cannot exceed the maximum primer distance.

Algorithm

Given the cDNA sequence and previous round information, the program first generates a possibly overlapping view of regions which follows logically from the last round results. As this view may contain overlapping regions with complicated structure, a transformation is made to generate a consistent non-overlapping view. A caution is that some information is lost in the process so as to make the regions non-overlapping. When the conservative approach is used, there is an extra possibility of refining - results as follows. If either primer defining the - result is involved in a + result, then the - result can be shrank to the primer midpoint. We reason as follows. Suppose that there is an exon boundary in the area between the outer primer end and the midpoint. Since the primer has a + result, it hybridizes regardless of the exon boundary within it. If this exon boundary is the only one that is involved in the - result, then the result should rather be + since the boundary should not affect the primer hybridization. It follows that there must be another boundary not between the outer primer end and the midpoint. Hence the shrinking is appropriate.

For each - or ? region in the non-overlapping view that has size larger than the new resolution, the program attempts to generate new primers within the region. However, no attempt is made to reuse primers within a region. A set of back-to-back primers are generated so that adjacent primers are not too close together (distance exceeding minimum distance between primer ends) while attempting to satisfy the new resolution. Thus, this strategy cannot honor requests for very small resolution. However, when there are adjacent + regions, very small resolution can be honored by generating a primer pointing to the + region and using an existing primer in the + region to form a primer pair.

There is a potential problem with the strategy of generating back-to-back primers within regions when the conservative approach is used. Since + results only span up to primer midpoints, the strategy leaves holes of ? regions when there are two adjacent + results. If the two + results are short, the problem can be solved easily by utilizing an existing primer pair that spans through the hole. Otherwise, the program does not generate new primers automatically to resolve this kind of very small ? regions, and manual primer design may be necessary. Finally, it is possible that no good primers can be found in a region. The program does not warn about this.

After new primers are generated, the program computes the set of all primer pairs that can possibly lead to improvements in refining regions. This set is usually larger than what one might want and the user is not obligated to use them all. Since only primer pairs that are close enough are generated, the user can reduce the maximum primer distance parameter to decrease the number of pairs that are automatically generated.

Output

- **Non-overlapping view**. This is the non-overlapping view that will be used for primer generation which shows all - and ? regions. Note that ? regions can be combined with an adjacent - region to form a larger - region.
- **Graphical view**. Shows a non-overlapping view of results in each round. Except for the last round, the view is derived from experimental results only. Special primer type or region specifications which add knowledge from other sources only show its effect in the last round. Directed triangles represent primers. White hollow triangles represent primers that only has ? results, yellow hollow triangles represent unused primers in a previous round that are useful in the new round, while yellow solid triangles represent newly generated primers. Otherwise, primers are white solid triangles. Red rectangles indicate + regions, orange rectangles indicate - regions,

while white rectangles indicate ? regions. Black vertical lines separate adjacent - regions. In test mode, known exon boundaries are shown by white vertical lines.

- **Plain file for input sequence**. This file can be saved (or be copied and pasted if there is problem with direct saving) in the user's local system for future reference. It is provided so that the user can conveniently get DNA sequences for primers.
- **Plain file for new round**. This file should be saved (or be copied and pasted) in the user's local system which can be modified to serve as input for the next round. It lists the current and the new sets of primers, primer pairs, and the possibly overlapping view of regions that follows logically from the results. Normally, the user should decide what experiments to carry out and fill out results for appropriate primer pairs. In test mode, since the gene structure is known, results are provided automatically assuming perfect experiments with no ? results. The software generates primer names in a mechanical way. There is no obligation for the user to name primers which are not automatically generated in the same way as long as the names fit the three-part rule described above.

Recommendations

It is recommended that the first round resolution be set to the expected length of exons (we used 150 in our sample of human genes) and the resolution be cut in half successively in later rounds (not exactly half, but with value slightly larger than half the previous value, to account for possible primer movements and the slight loss of precision in + results when the conservative approach is used). Alternatively, when the cDNA is very long or when a lot of exon boundaries are expected, it is reasonable to start with a relatively large first round resolution. In later rounds, if we find that the progress in region refinements is slow, we can speed up the decrease of resolution. On the other hand, if the progress is fast, we can slow down the decrease of resolution to save primers.

The user is free to add additional information to the last round input which might not be the results of PCR experiments as long as the input format is obeyed. For example, the user can specify the optional primer type when defining a primer (to force the program to believe that the primer is involved in a result with the specified strongest type without really having a corresponding result) or specify the type of a region without experimental support.

Results of the primer generation algorithm are meant to be suggestive. Sometimes, generated primers can be obviously bad to the eye due to the complicated nature of the algorithm. User is free to delete unused primer pairs, move primers, correct or ignore inconsistent results. But, please remember to make a lot of backups of the new round file. Run as many times as you like to decide on appropriate input parameters. The current status can always be checked by simply taking the new round file and run the program again with it as the last round input without changing the file or decreasing the resolution. Keep in mind that the non-overlapping and the graphical views do not provide the most information about the current status. The new round file contains the most general possibly overlapping view.

Appendix I. Methods, Tools and Databases

The GRAIL Gene Recognition System

GRAIL is a modular system which supports the recognition of gene features and gene modeling for the analysis and characterization of DNA sequences. GRAIL uses multiple hybrid statistical and neural network-based pattern recognizers and a dynamic programming approach to constructing gene models. GRAIL recognizes protein coding regions (exons), poly-A addition sites, potential promoters, CpG islands and repetitive DNA elements. XGRAIL also has a direct link to the genQuest server allowing characterization of newly obtained sequences by homology based methods through accessing a number of databases using a number of comparison methods including: FASTA, BLAST and a parallel implementation of the Smith-Waterman algorithm which utilizes the DEC cluster at ORNL CSMD. Following an analysis session the user can use an annotation tool to generate a ``feature table'' describing the current sequence and it properties. All of this information is presented to the user in graphic form in the X-window based client-server system XGRAIL.

Since its development in 1991 (Uberbacher and Mural, 1991), the GRAIL system at ORNL has become the world standard for predicting protein coding regions (exons) and modeling genes in DNA sequences. From its inception GRAIL has been accessible over the network in a variety of ways including e-mail, a X-based client-server system and through various web browsers over the world wide web. The experience with GRAIL at ORNL gives us an understanding of the needs of the genomic / biomedical research community. The GRAIL system currently analyzes about 17 million bases of DNA sequence per month (using methods which are simpler and less comprehensive than what is proposed in this GC project). In addition the ORNL Informatics Group maintains a public server, genQuest, which allows investigators to query a number of public sequence databases to establish whether a newly determined

- Solovyev V.V., Kell A.E., Rogozin I.B., Kolchanov N.A. The application of computers in Molecular Biology. The theory of genetic texts. Novosibirsk State University (Russ),90 p.,(1987).
- Kolchanov N.A.,Solovyev V.V.,Shindyalov I.N.,Strelets V.B. The application of computers in Molecular Biology. The structural organization of globular proteins. Novosibirsk State University (Russ),(1986),90 p.
- Kell A.E., Kolchanov N.A.,Solovyev V.V.The mathematical modeling in the molecular biology and genetics. The theory of the mutational process:deletions and duplications. Novosibirsk State University (1989),90 p.

Articles:

- Solovyev V.V., Salamov A.A. (1999) INFOGENE: a database of known gene structures and predicted genes and proteins in sequences of genome sequencing projects. **Nucl.Acid Res.**, 27,248-250.
- Alexandrov N.N., Solovyev V.V.(1998) Statistical significance of ungapped sequence alignment. In Pacific Symposium on Biocomputing'98 (eds. Altman R.,Dunker K.,Hunter L., Klein T.), p.463-472.
- Kolchanov NA, Ponomarenko MP, Kel AE, Kondrakhin YuV, Frolov AS, Kolpakov FA, Goryachkovsky TN, Kel OV, Ananko EA, Ignatieva EV, Podkolodnaya OA, Babenko VN, Stepanenko IL, Romashchenko AG, Merkulova TI, Vorobiev DG, Lavryushev SV, Ponomarenko YuV, Kochetov AV, Kolesov GB, Solovyev VV, Milanesi L, Podkolodny NL, Wingender E, Heinemeyer T.(1998) GeneExpress: a computer system for description, analysis, and recognition of regulatory sequences in eukaryotic genome.In Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (eds.Gaasterland T. et al.),95-104
- Agulnik A.I.,Bishop C.E., Lerner J.L., Agulnik S.I., Solovyev V.V.(1977) Analysis of mutation rate in the SMCY/SMCX genes shows that mammalian evolution is male driven. **Mammalian Genome** 8, 134-138.
- Salamov A.A., Solovyev V.V. (1997) Protein secondary sturcture prediction using local alignments. **J. Mol.Biol**,268,1, 31-36..
- Solovyev V.V., Salamov A.A. (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (eds.Rawling C.,Clark D.,Altman R.,Hunter L., Lengauer T.,Wodak S.), Halkidiki, Greece, AAAI Press,294-302.
- Salamov A.A., Solovyev V.V. (1997) Recognition of 3'-end cleavage and polyadenilation region of human mRNA precursors. **CABIOS** 13, 1, 23-28.
- Alexandrov N., Solovyev V. (1996) Effect of Secondary Structure Prediction on Protein Fold Recognition and Database Search. The Seventh Workshop on Genome Informatics (GIW96), Tokyo, Japan.
- Kelly R.L., Solovyeva I., Lyman L.M., Richman R., Solovyev V.V., Kuroda M. (1995) Expression of MSL-2 causes assembly of dosage compensation regulators on the X chromosomes and female lethality in Drosophila. **Cell**,81,867-877.
- Solovyev V.V., Salamov A.A., Lawrence C.B. (1995) Prediction of human gene structure using linear discriminant functions and dynamic programming. In Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (eds. Rawling C.,Clark D.,Altman R.,Hunter L.,Lengauer T.,Wodak S.), Cambridge,England, AAAI Press,367-375.
- Salamov A.A., Solovyev V.V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. **J. Mol.Biol.** (1995), 247,1,11-15.
- Solovyev V.V., Salamov A.A. Predicting a-helix and b-strand segments of globular proteins. **CABIOS** (1994),v.10,6,661-669.
- Solovyev V.V., Salamov A.A., Lawrence C.B. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. **Nucl. Acids Res.** (1994) 22,N 24,5156-5163.
- Solovyev V.V., Salamov A.A., Lawrence C.B. 1994. The prediction of human exons by

sequence by using the appropriate group of weight matrices. The group matrix for vertebrates genomes is used as default. It is also possibile to choose the matrices to be used individually.

15. Mail to (Yes, No).

Results of GeneBuilder in text format can be sent by e_mail. This is useful when the analysis is of long sequences or is over poor network connections. In this case the e_mail address is mandatory.

Availibility

WEBGENE

References

- Milanesi L., D'Angelo D., Rogozin I.B. GeneBuilder: interactive in silico prediction of genes structure. Bioinformatics, 1999, Jul; 15 (7):612-621.

- Milanesi L. and Rogozin I.B. Prediction of human gene structure. In: Guide to Human Genome Computing (2nd ed.) (Ed. M.J.Bishop) Academic Press, Cambridge, 1998, 215-259.

Close this window

4: Prediction of complete gene structures in human genomic DNA - Burge, Karlin - 1997
4: Predicting internal exons by oligonucleotide composition and discriminant analys.. (context) - Solovyev, Salamov et al. - 1994
4: A generalized hidden Markov model for the recognition of human genes in DNA - Kulp, Haussler et al. - 1996

BibTeX entry: (Update)

Salzberg, S. (1997). A method for identifying splice sites and translational start sites in eukaryotic mRNA. Computer Applications in the Biosciences (CABIOS). to appear. http://citeseer.nj.nec.com/salzberg97method.html More

```
@misc{ salzberg97method,
  author = "S. Salzberg",
  title = "A method for identifying splice sites and translational start sites in
    eukaryotic mRNA",
  text = "Salzberg, S. (1997). A method for identifying splice sites and translationa
    start sites in eukaryotic mRNA. Computer Applications in the Biosciences
    (CABIOS). to appear.",
  year = "1997",
  url = "citeseer.nj.nec.com/salzberg97method.html" }
```

Citations (may not include all citations):
24 Identification of coding regions in genomic DNA sequences: A.. (context) - Snyder, Stormo - 1993
24 Prediction of gene structure (context) - Guigo, Knudsen et al. - 1992
22 Predicting internal exons by oligonucleotide composition and.. (context) - Solovyev, Salamov et al. - 1994
20 Prediction of human mRNA donor and acceptor sites from the D.. - Brunak, Engelbrecht et al. - 1991
10 Improved splice site detection in genie - Reese, Eeckman et al. - 1997
10 A weight array method for splicing signal analysis (context) - Zhang, Marr - 1993
9 Finding genes in human DNA with a hidden Markov model - Henderson, Salzberg et al. - 1997
9 The gene identification problem: an overview for developers (context) - Fickett - 1996
8 Grail: A multi-agent neural network system for gene identifi.. (context) - Xu, Mural et al. - 1996
5 noncoding sequences from 699 vertebrate messenger RNAs (context) - Kozak, of - 1987
5 and applications to genome project (context) - Senapathy, Shapiro et al. - 1990
5 Identification of coding regions in genomic DNA (context) - Snyder, Stormo - 1995
4 Finding genes in DNA using decision trees and dynamic progra.. (context) - Salzberg, Chen et al. - 1996
3 Splicing signals in drosophila: intron size (context) - Mount, Burks et al. - 1992
3 A consideration of alternative models for the initiation of .. (context) - Kozak - 1992
1 AC introns: An ATtACk on dogma (context) - Mount - 1996

Documents on the same site (http://www.cs.jhu.edu/~salzberg/home.html): More
Locating Protein Coding Regions in Human DNA using a Decision.. - Salzberg (1995) (Correct)
Finding Genes in DNA with a Hidden Markov Model - Henderson, Salzberg, Fasman (1997) (Correct)
A Weighted Nearest Neighbor Algorithm for Learning with.. - Cost, Salzberg (1993) (Correct)

A Method for Identifying Splice Sites and Translational Start Sites in Eukaryotic mRNA (1997) (Correct) (8 citations)

Steven L. Salzberg

Abstract: This paper describes a new method for determining the consensus sequences that signal the start of translation and the boundaries between exons and introns (donor and acceptor sites) in eukaryotic mRNA. The method takes into account the dependencies between adjacent bases, in contrast to the usual technique of considering each position independently. When coupled with a dynamic program to compute the most likely sequence, new consensus sequences emerge. The consensus sequence information is summarized in conditional probability matrices which, when used to locate signals in uncharacterized genomic DNA, have greater sensitivity and specificity than conventional matrices. Species-specific... (Update)

Context of citations to this paper: More

...have primarily classi ed sequences according to the marginal probabilities of nucleotides occuring in positions around the sites. **Salzberg [12] analyzed both start sites and splice junctions by assuming that the probability of a nucleotide appearing in a position p depended on...**

.... **position speci#c nucleotide distributions #Staden 1984#, correlations between positions #Agarwal Bafna 1998; Burge Karlin 1997; Salzberg 1997; Zhang Marr 1993#, and higher order e#ects using machine learning techniques #Pedersen Nielsen 1997#.** In this paper, we make...

Cited by: More

Evidence of Limited Structural Organization in Globin Intron .. - Dawson, Yamamoto (Correct)
Mining Biomolecular Data Using Background Knowledge and.. - Qicheng Ma Jason (Correct)
The Conserved Exon Method for Gene Finding - Bafna, Huson (2000) (Correct)

Active bibliography (related documents): More All

1.1: Modeling DNA Splice Regions by Learning Bayesian Networks - Dash, Gopalakrishnan (2001) (Correct)
0.9: A Decision Tree System for Finding Genes in DNA - Salzberg, Delcher, Fasman.. (1997) (Correct)
0.8: Finding Genes in DNA with a Hidden Markov Model - Henderson, Salzberg, Fasman (1997) (Correct)

Users who viewed this document also viewed: More All

0.3: Prediction of human mRNA donor and acceptor sites from .. - Brunak, Engelbrecht.. (1991) (Correct)
0.3: Computational Genefinding - Haussler (1998) (Correct)
0.3: A Generalized Hidden Markov Model for the Recognition .. - Kulp, Haussler.. (1996) (Correct)

Similar documents based on text: More All

0.9: Human GC-AG alternative intron isoforms with weak donor sites .. - Thanaraj, Clark (2001) (Correct)
0.8: Clustering and Detection of 5' Splice Sites of mRNA by k.. - Murakami, Takagi (Correct)
0.8: Clustering and Detection of 5' Splice Sites of mRNA by - Weight Matrix Model (Correct)

Related documents from co-citation: More All